Early stopping for kernel boosting algorithms

Yuting Wei*, Fanny Yang*, Martin Wainwright

Department of Statistics and EECS, University of California, Berkeley

Problem Setting

- Given **arbitrary regular loss function** $\phi : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, n fixed covariates x_i and corresponding random $Y_i \sim \mathbb{P}_{x_i}$
- Object of interest: minimizer of population loss function over some **function class $\mathcal{F}$**

$$\mathcal{L}(f) := \mathbb{E}_{Y_1^n}\Big[\frac{1}{n}\sum_{i=1}^{n}\phi(Y_i, f(x_i))\Big] \quad \text{and} \quad f^* := \arg\min_{f\in\mathcal{F}} \mathcal{L}(f)$$

- In practice: minimizer of empirical loss function based on observed $\{x_i, Y_i\}_{i=1}^n$

$$\mathcal{L}_n(f) := \frac{1}{n}\sum_{i=1}^{n}\phi(Y_i, f(x_i)) \quad \text{and} \quad \widehat{f} := \arg\min_{f\in\mathcal{F}} \mathcal{L}_n(f)$$

- If function class large $\mathcal{F} \to$ risk of **overfitting** to noise!
- Standard way to prevent overfitting: additive penalty function

Boosting algorithms

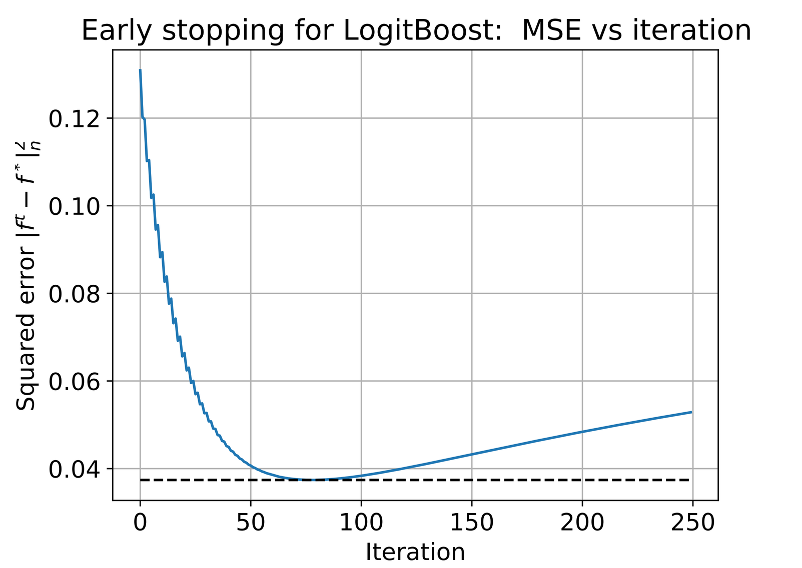
- Based on a sequence of additive updates (weak learners) to improve the fit of a function, see e.g. [1]
- Can be viewed as functional gradient descent steps with updates

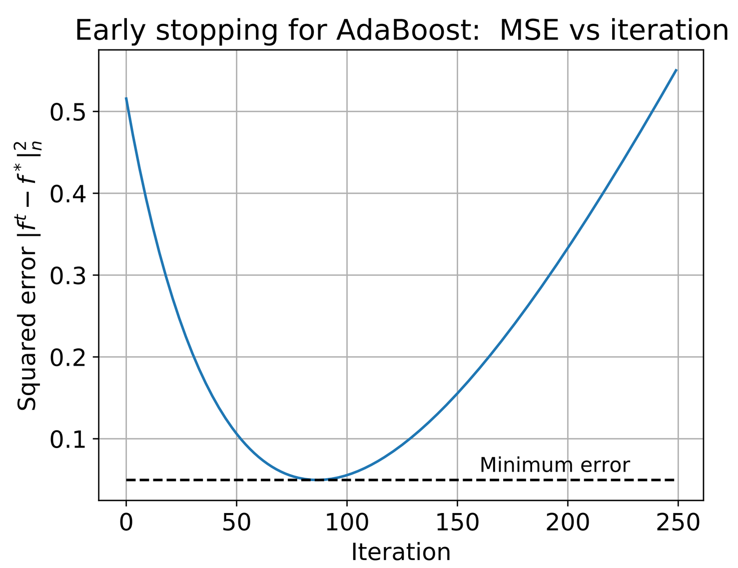
$$f^{t+1} = f^t - \alpha^t g^t \quad \text{with} \quad g^t \propto \arg\max_{\|d\|_{\mathcal{F}}\le 1} \langle \nabla\mathcal{L}_n(f^t),\, d(x_1^n)\rangle \qquad (1)$$

Kernel Boosting

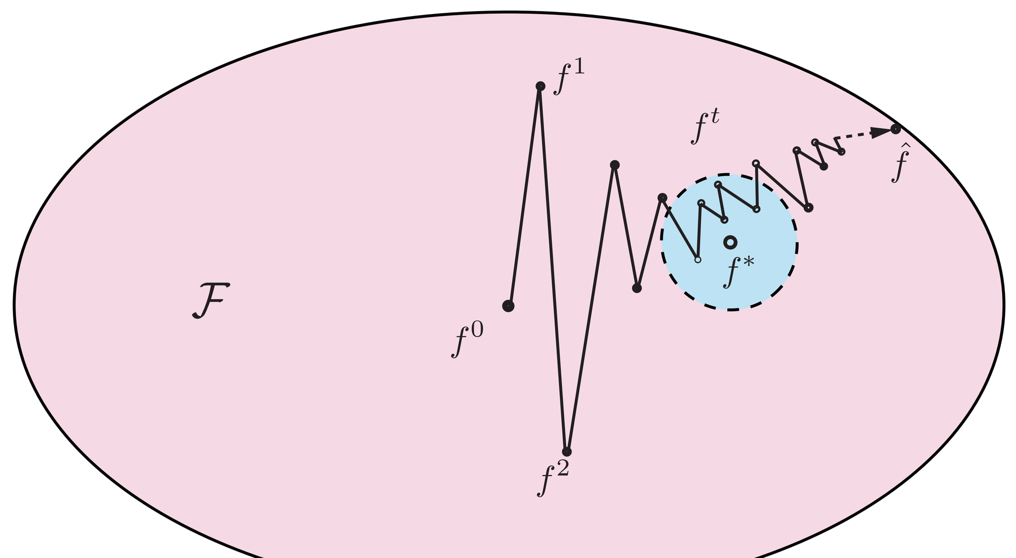
- A positive semidefinite *kernel function* $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
  e.g. Gaussian $\mathbb{K}(x,z) = \mathrm{e}^{\frac{(x-z)^2}{2\sigma^2}}$, Sobolev $\mathbb{K}(x,z) = 1 + \min\{x,z\}$
  induces a **Reproducing Kernel Hilbert Space** (RKHS) $\mathcal{H}$
- Optimal $\widehat{f} \in \mathcal{H}$ can be represented as $f(\cdot) = \sum_{i=1}^n \omega_i \mathbb{K}(\cdot, x_i)$
- Define kernel matrix K on covariates $\{x_i\}_{i=1}^n$ by $K_{ij} = \mathbb{K}(x_i, x_j)$
- **Kernel boosting update** for $\mathcal{F} = \mathcal{H}$ in (1) on vectors

$$f^{t+1}(x_1^n) = f^t(x_1^n) - \alpha n K \nabla\mathcal{L}_n(f^t) \qquad (2)$$





Overfitting and early stopping



Running until convergence may overfit $\to$ stop early!

For **least-squares loss**, early stopped boosting (*algorithmic regular.*) and penalized estimators behave similarly statistically, i.e.

$$\|f_{\text{pen}} - f^*\|_n^2 \sim \|f^T - f^*\|_n^2$$

for an appropriate stopping time T, see e.g. [2,3,4].

Main contributions

- Is there a common principle behind statistical behavior of algorithmic and penalized regularization?

  YES! We prove statistical rates for early stopping using ***the same key quantities as in penalized regularization***

- Can we extend to other loss functions?

  YES! Our new proof technique allows to extend to a ***broad class of loss functions (e.g. AdaBoost, LogitBoost ...)***

Key quantities

- Key quantity I: **Localized Gaussian complexity**
  $\mathcal{G}_n(\mathcal{E}(\delta,1)) := \mathbb{E}\Big[\sup_{g\in\mathcal{E}(\delta,1)} \frac{1}{n}\sum_{i=1}^n w_i g(x_i)\Big], \quad w_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$
  where: $\mathcal{E}(\delta,1) := \Big\{f - g \mid f, g \in \mathcal{H}, \|f-g\|_{\mathcal{H}} \le 1, \|f-g\|_n \le \delta\Big\}$
- Key quantity II: **Critical radius** δ_n smallest scalar that satisfies

$$\frac{\mathcal{G}_n(\mathcal{E}(\delta,1))}{\delta} \le \frac{\delta}{\sigma}$$

Main results

Theorem 1. *Given some regular loss function ϕ and the function iterates $\{f^t\}_{t=0}^{\infty}$ as in (2), for all iterations $T = 0, 1, \ldots \lfloor 1/(8\delta_n^2)\rfloor$, the averaged function estimate $\bar{f}^T$ satisfies with high probability*

$$\mathcal{L}(\bar{f}^T) - \mathcal{L}(f^*) \le C\Big(\frac{1}{\alpha T} + \delta_n^2\Big), \quad \text{and}$$
$$\|\bar{f}^T - f^*\|_n^2 \le C\Big(\frac{1}{\alpha T} + \delta_n^2\Big).$$

Examples for specific kernel spaces:

- **γ-exponential decay**: the kernel eigenvalues $\mu_j \le c_1 \exp(-c_2 j^\gamma)$, when stopped after $T \asymp \frac{n}{\log^{1/\gamma} n}$ steps:

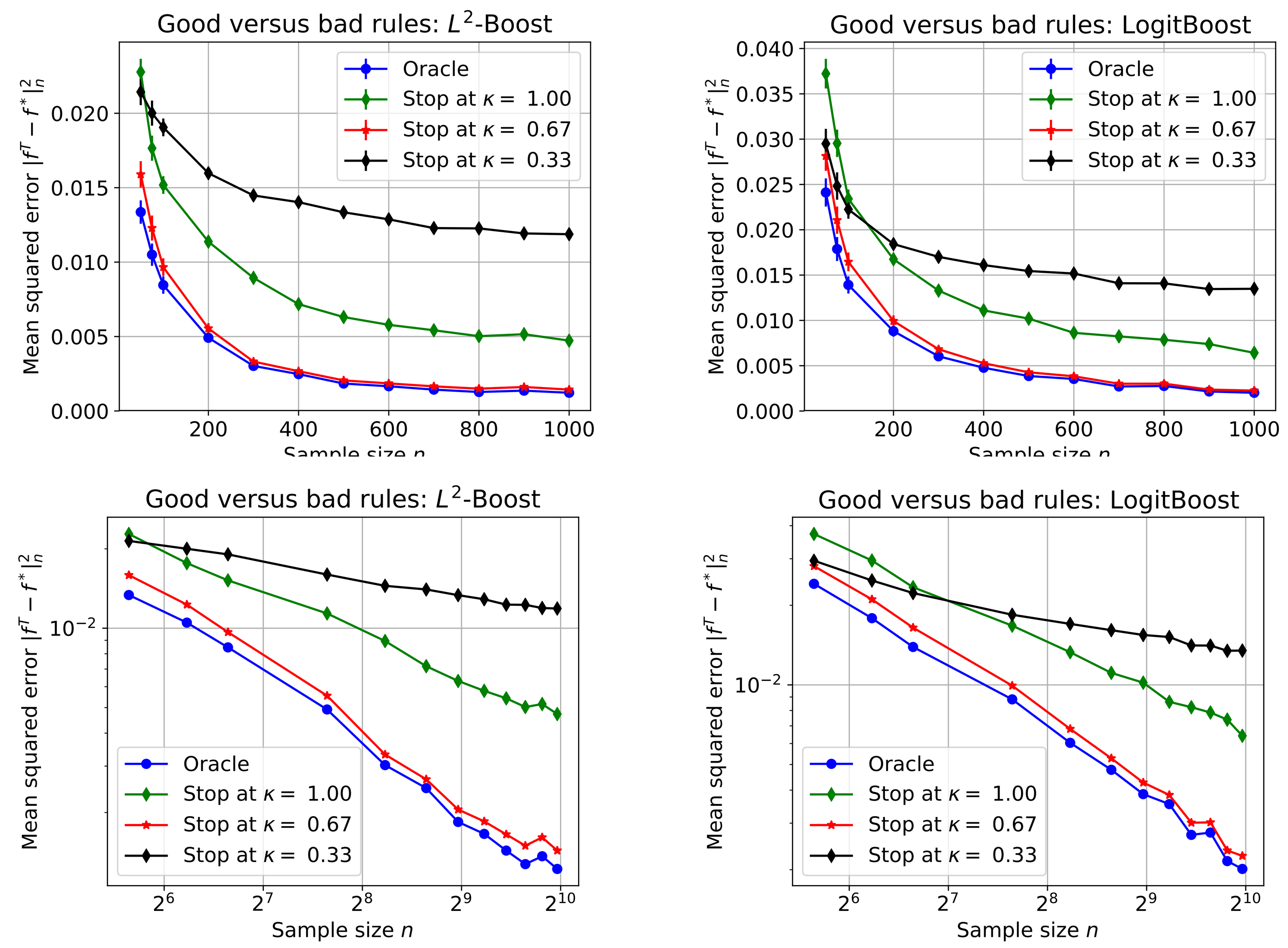
$$\|\bar{f}^T - f^*\|_n^2 \lesssim \frac{\log^{1/\gamma} n}{n}$$

- **β-polynomial decay**: the kernel eigenvalues $\mu_j \le c_1 j^{-2\beta}$, when stopped after $T \asymp n^{2\beta/(2\beta+1)}$ steps:

$$\|\bar{f}^T - f^*\|_n^2 \lesssim n^{\frac{-2\beta}{2\beta+1}}$$

Numerical results

$\mathcal{H}$: first order Sobolev space, stop iterates after $T = (2n)^\kappa$



References

[1] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Nonlinear estimation and classification*. Springer, 2003, pp. 149–171.
[2] P. Bühlmann and B. Yu, "Boosting with L^2 loss: Regression and classification," *Journal of American Statistical Association*, vol. 98, pp. 324–340, 2003.
[3] Y. Yao, L. Rosasco, and A. Caponnetto, "On early stopping in gradient descent learning," *Constructive Approximation*, vol. 26, no. 2, pp. 289–315, 2007.
[4] G. Raskutti, M. Wainwright, and B. Yu, "Early stopping and non-parametric regression: An optimal data-dependent stopping rule," *Journal of Machine Learning Research*.